# EFFECTIVENESS OF CERTAIN EXPERIMENTAL PLANS UTILIZED IN SENSORY EVALUATIONS

J. Wayne Hamman and Jan Eindhoven
Armed Forces Food and Container Institute
Chicago, Illinois

First, I would like to present some specific purposes of sensory testing at the Armed Forces Food and Container Institute. This will be followed by a discussion of the experimental results obtained from the sensory evaluation of four meat products.

SOME PURPOSES OF SENSORY TESTING. You are aware of the numerous food items developed, purchased, stored and consumed by the Armed Forces. A continuous program exists at the Institute to determine whether or not differences in quality or stability exist between different samples of food. Here are some of the most common requirements for conducting these sensory tests:

1. Pre-award evaluation for intent to purchase. When a certain food item (such as peanut butter) is required by the Army, it advertises for bids from manufacturers. Those manufacturers who are interested submit samples of their products for preference evaluation. These samples are taste-tested, and those that are reliably poorer than our standard products are rejected. In this way, sensory testing screens out lower quality products that are relatively unacceptable to the soldier-consumer.

2. Storage stability. Since foods may not be used for several years after they are packed, a considerable amount of research is devoted to extend the shelf-life of a food. Sensory tests are concerned with the preference or intensity of off-flavor changes that take place during storage. Sensory tests are made on foods stored at different temperatures over time up to two years, and more.

3. Packaging studies. Often, a flexible package may be desired for use in the field. However, the relative storage life of food in such a plastic package must be considered if a change is to be made from a canned food.

4. Processing variables. New processing and preservation methods of foods, such as freeze-dehydration of meats, offer new problems in flavor and texture for evaluation. It must be determined whether or not this new product is as desirable as the existing food prepared by other methods.

5. Special Purpose Foods. Recent evaluations have included novel preparations of foods designed specifically for space flight. For example, meat dishes that may be consumed through a straw. Also the new Quick-Serve Meals have been developed which consist largely of pre-cooked dehydrated foods. Preliminary testing is done first at the Institute to determine whether or not these foods are satisfactory enough for further testing in the field among astronauts and soldiers.

SENSORY EVALUATION LABORATORY. In the sensory evaluation laboratory careful attention is given to assure that each sample of food is treated in the same way as every other one in an evaluation. Some of the procedures followed include

1. The random assignment of code numbers to the samples so that subjects will not be biased.

2. In order for the individual to regain sensitivity, that is to get the flavors of a previous sample out of his system, a 30-second time interval is specified between the time that a subject returns the rating of his previous sample and when he receives his next one. Automatic timers are used.

3. In a sensory test each sample is served first, second, third, etc., an equal number of times to minimize position effect. When the number of subjects permits, all possible serving sequences of samples are used. to reduce both serving position and sequence biases that might exist.

4. The number of samples that a subject receives is normally limited to four in order to minimize effects of fatigue and to maintain interest in the evaluation.

PURPOSE OF EXPERIMENT. This experiment is concerned with the effect on sensory results when meat samples are presented to subjects in different combinations and sequences. Specific topics considered are

1. Sequence effects. How is the rating of a sample influenced by the quality of a preceding sample? It is hypothesized that when more highly preferred samples precede those of relatively low preference, the difference between them is emphasized. It is further hypothesized that when the more highly preferred samples follow those of relatively low preference, the difference between them is reduced.

2. Position effects. How is the rating of a sample influenced by the number of preceding samples he has evaluated?

3. Matnitude of error term. How is the size of the error term affected by the quality and sequences of the samples presented?

MATERIALS AND METHODS. Four meat products were evaluated in this experiment: ham, pork, chicken (white) and chicken (dark). Four samples of each of the meat products consisted of two control and two treated samples. An additional preparation variable was included for each meat product which causes both the two control samples and the two treated samples to be considered as non-duplicates. However, the determination of the effect of the additional variable is not the intent of this experiment and will not be specifically considered in this paper.

The subjects sat in a semi-enclosed testing booth for privacy in making evaluations. Each individual received four samples of one of these meat products. These samples were presented one at a time through a turn-table in a wall separating the booth from the serving and preparation area. The subject was asked to state his preference for each sample on a nine-point rating scale. The terminology on this hedonic scale[5] ranged from dislike extremely, coded 1, to like extremely, coded 9, and is shown on the illustrated EAM card (Figure 1) [Figures and Tables can be found at the end of this article] which is used for rating and mechanical data reduction.

Subjects were selected at random from a pool of about 450 employees.

EXPERIMENTAL PLANS. Five experimental plans are considered:

(1) 4! : Conventional plan with all sequences of serving orders. Twenty-four subjects are required for this plan in order to encompass all sequences.

(2) cccc: Two control samples balanced over four serving positions. The two control samples, you will recall, are the non-treated samples and differ by a preparation variable.

(3) tttt: Two treated samples balanced over four serving positions.

(4) cctt: Two control samples followed by two treated samples. The two control samples were served equally often in positions 1 and 2; the two corresponding treated samples were served equally often in positions 3 and 4.

(5) ttcc: Two treated samples followed by two control samples. The two treated samples were served equally often in positions 1 and 2; the two corresponding control samples were served equally often in positions 3 and 4.

Twenty-four subjects were selected at random for each of these plans; all plans were carried out for each of the four products.

EXPERIMENTAL DESIGN. A latin square experimental design[1] was utilized in this study in order to

a. determine the effect, if any, of the sequence of the presentation on the rating for a sample, and

b. reduce the experimental error, if a position effect was present.

Six replications of a 4x4 latin square design were utilized in each plan.

Figure 2 illustrates the allocation of samples for the 4! conventional plan which has all possible sequences. Subjects 1, 2, 3, 4, constitute the first replication, then, and the four samples A, B, C, D, all occur once in each order in a replicate and a subject receives all four samples. The basic Analysis of Variance components, before isolating certain individual degrees of freedom is also given as a part of Figure 2. Since the subjects in a replicate were selected at random, no difference was anticipated between replicates. In the Analysis of Variance treatment x replication and order x replication interactions were pooled into the error term, since there is no reason to expect that these interactions are real.

SEQUENCE RESULTS. Results for the conventional plan (4!) which had all serving orders are shown in Table 1. Mean preference scores for controls were higher than treated for all four of the meat products. The mean differences ranged from 0.56 to 0.79 scale points and significance values were no larger than P = .06 in testing the null hypothesis, namely, that the control mean and treated mean are the same.

Table 2, which presents results for the plan with two control samples served first followed by two treated samples, shows a substantial increase in the discrimination between control and treated samples. The

combined mean difference increased from 0.69 for the conventional plan to 1.00 for this cctt plan and individual probability values declined with the exception of pork which remained about the same (P = .01 vs .02). This phenomenon has been described in previous studies.

In a study with soups and beverages [4] the situation of poor samples following good ones was termed "contrast", that is, one of emphasizing differences. An explanation hypothesized for the phenomenon of "contrast" was that the positive qualities of the good sample are either noticed to be absent or bad qualities are noticed as present in the poor one, thus emphasizing in either case the short-comings of the less preferred one.

Results of the alternative situation where relatively poor samples precede the good ones are given in Table 3. You will notice that the direction of the differences determined in the conventional 4! and contrast plans are not found here. The combined treated samples were rated 0.06 scale points higher than the control in this ttcc plan and none of the individual product differences were statistically significant. This situation where poor samples preceded good ones was termed "convergence"[4], although with these liquids, convergence effects were not found to be statistically significant. An explanation hypothesized is that the presentation of a "poor" sample increases an individual's awareness of the presence of some of the negative characteristics in a "good" sample[4].

Conclusions drawn from these results might be modified somewhat, after a consideration of the position of presentation. We might ask the question: what part of the observed contrast and convergence effects might be due to the fact that samples were presented in positions 3 and 4? We will now proceed to an examination of the positional effect of presentation.

POSITIONAL RESULTS. An examination of the effect of the order in which the sample was received on its rating has been made, considering all five plans. Mean scores by position are given in Table 4. Combined positional means are given in the lower part of this Table for each type of plan. Hypothesis tested were

$$H_1 : \bar{x}_1 > \bar{x}_2$$

$$H_2 : \bar{x}_3 > \bar{x}_4$$

$$H_3 : \bar{x}_{1+2} > \bar{x}_{3+4}$$

Since it was theorized that the position effect (if one existed) may be dependent upon the quality of the product, data were analyzed separately for the control (cc), treated (tt) and the mixed sequence (ct or tc). These are the breakdowns, then, in Table 5. The orthogonal comparisons are shown here for positions 1 vs 2, 3 vs 4 and 1+2 vs 3+4, relating to the mean scores of Table 4.

Probabilistic results from these similar experiments of ham, pork, chicken (white), and chicken (dark) were combined[3] in order to strengthen evidence concerning an effect of position.

Combined evidence for the control (cc) type pairing demonstrated a decrease in preference from position 3 to 4 ($P = .06$) and from the first two positions to the second two positions ($P = .05$). Evidence was not conclusive concerning the decrease in preference exibited from position 1 to 2 ($P = .12$). In the latter case the decrease was 0.22 scale points and was in the hypothesized direction.

Combined evidence for the treated (tt) type pairing did not demonstrate significant positional effects. In the all treated plan the mean for the first two positions was 6.74 contrasted with 6.70 for the last two positions which is reflected by chi-square[6] probability of 0.48. Also the differences between mean preferences, regarding positions 1 vs 2 and 3 vs 4, were not statistically significant.

In the mixed pairing (ct) position 1 was shown to be significantly higher than position 2 ($P = .03$) and the first two positions significantly higher than the last two positions ($P = .03$). Little difference was evidenced between means for positions 3 and 4, however.

A taste testing experiment was reported previously[7] which studied the effect of fatigue over a series of eight samples presented in one sitting. The two foods considered were canned sauerkraut and canned bread with margarine. In a comparison of serving positions 1 or 3 with 5 or 8,

there was found to be no significant difference in preference rating due to the position of the test food, whether it appeared early or late in the eight sample series. These results on the treated samples (no significant difference between positions) would seem to bear out those found on sauerkraut and margarine, since these food items are all relatively low preference items, particularly, the margarine of ten years ago.

Credulence is lent to the theory, then, that while a position effect does appear to exist, the quality of the sample determines whether or not there is a decline in preference with the sequence of presentation.

MAGNITUDE OF ERROR TERMS. A comparison of the magnitude of error terms for the different plans is presented in Table 6. In an analysis of Variance of these variances followed by a Duncan Multiple Range[2] test of means it was determined that the variance for the all control plan was significantly ($P = .05$) smaller than the three plans having both control and treated samples. Also, the error variance of the all treated plan was significantly ($P = .05$) smaller than for the conventional 4! and the plan with two control samples followed by two treated samples (cctt).

These results are in line with anticipations, however, since the ranges of ratings of individuals within the all treated and all control plans are less than for the other plans. Hence, the magnitude of disagreement in preference would be expected to be less for these plans. The differences in magnitude of these variances, though, do point out the necessity for analyzing differences between means, such as those for order, individually, for each plan.

SUMMARY. The preference rating scale is normally used for comparative purposes between samples. However, one can see the effect that might occur on the magnitude of scores of experimental samples, depending upon the quality of standards or controls with which they are compared.

Also the sequence in which relatively good and poor samples were presented made a considerable difference in what conclusions would be drawn. Evidence concerning the effect of the position of presentation was given. For the higher preference samples it was demonstrated that a fatigue effect or sensitivity effect was present causing a decline in the rating of subsequent samples (although not so conclusive between positions 1 and 2 where $P = .12$). For the lower preference samples there was not determined to be a decline in ratings in subsequent positions.

If one desired to minimize Type II statistical error (acceptance of the null hypothesis when it is false), then, one would wish to present test samples after the standards. If the test samples were poorer, the difference would be emphasized by a contrast effect pointed out earlier. The statistical soundness of such a procedure might be questioned, however, since the magnitude of the contrast effect would be expected to be greater when the difference between samples was greater. This would affect the probability statements concerning a "true" difference in the population to a corresponding unknown degree.

## REFERENCES

(1) Cochran, G. C. and Cox, G. M. Experimental Designs, John Wiley and Sons, New York, 1950, pp. 103-112.

(2) Duncan, David B. Multiple Range and Multiple F Tests, Biometrics, March 1955, pp. 1-42.

(3) Jones, L. V. and Fiske, D. W. Models for Testing the Significance of Combined Results, The Psychological Bulletin, Vol. 50, No. 5, pp. 375-382.

(4) Kamenetsky, Joe, Contrast and Convergence Effects in Ratings of Foods, Journal of Applied Psychology, Vol. 43, No. 1, 1959.

(5) Peryam, David R., et. al. Food Preferences of Men in the U. S. Armed Forces. Department of the Army, Quartermaster Food and Container Institute for the Armed Forces, January 1960.

(6) Snedcor, G. W. Statistical Methods, The Iowa State College Press, Ames, Iowa, 1946.

(7) Symposium sponsored by the Quartermaster Food and Container Institute for the Armed Forces, Food Acceptance Testing Methodology, October 1953, pp. 92-99.

**Figure 1. Preference Rating Card Used for Sensory Evaluations**

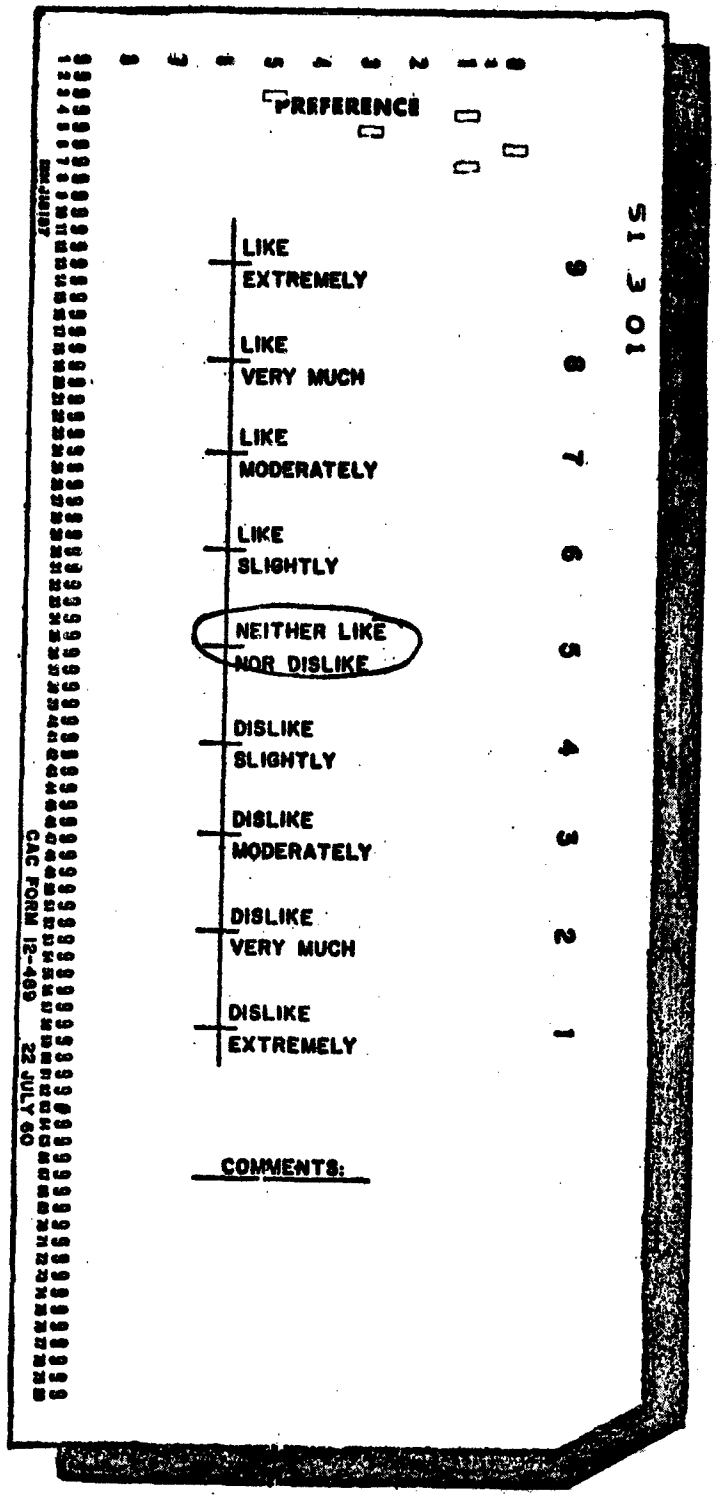

FIGURE 2. Allocation of Samples for the 4! Experimental Plan Utilizing a Latin Square Design

| Subject | Order 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1, 5, 9, 13, 17, 21 | A | B | C | D |
| 2, 6, 10, 14, 18, 22 | D | A | B | C |
| 3, 7, 11, 15, 19, 23 | C | D | A | B |
| 4, 8, 12, 16, 20, 24 | B | C | D | A |

Analysis of Variance Componenets

| Source of Variation | Degrees of Freedom |
|---|---|
| Subjects | 23 |
| Orders | 3 |
| Treatments | 3 |
| Error | 66 |
| Total | 95 |

TABLE 1. Mean Preference Ratings of Four Meat Samples in All Possible Serving Orders, Termed a 4! Plan*

| Food | Control | Treated | Control - Treated Diff. | Signif. Level for Diff. |
|---|---|---|---|---|
| Ham | 7.17 | 6.46 | 0.71 | .04 |
| Pork | 6.56 | 5.77 | 0.79 | .01 |
| Chicken, W. | 7.33 | 6.77 | 0.56 | .02 |
| Chicken, D. | 6.75 | 6.06 | 0.69 | .06 |
| Combined | | | 0.69 | |

TABLE 2. Mean Preference Ratings of Four Meat Samples Where Two Control Samples Were Presented, Followed By Two Treated Samples, Termed a cctt Plan*

| Food | 1st & 2nd Position Control | 3rd & 4th Position Treated | Control - Treated Diff. | Signif. Level for Diff. |
|---|---|---|---|---|
| Ham | 7.40 | 5.75 | 1.65 | .001 |
| Pork | 6.52 | 5.88 | 0.64 | .02 |
| Chicken, W. | 7.14 | 6.35 | 0.79 | .005 |
| Chicken, D. | 6.71 | 5.81 | 0.90 | .005 |
| Combined | | | 1.00 | |

TABLE 3. Mean Preference Ratings of Four Meat Samples Where Two Treated Samples Were Presented, Followed By Two Control Samples, Termed an ttcc Plan*

| Food | 1st & 2nd Position Treated | 3rd & 4th Position Control | Control - Treated Diff. | Signif. Level for Diff. |
|---|---|---|---|---|
| Ham | 6.85 | 6.60 | -0.25 | n.s. |
| Pork | 5.96 | 5.92 | -0.04 | n.s. |
| Chicken, W. | 6.90 | 7.04 | 0.14 | n.s. |
| Chicken, D. | 5.90 | 5.83 | -0.07 | n.s. |
| Combined | | | -0.06 | |

*Individual Means represent ratings of 2 samples by 24 subjects.

TABLE 4. Mean Preference Ratings for Each Serving Order of Four Meat Products for Five Experimental Plans. Individual Means Represent 24 Subjects.

| Food | Plan/ | Order of Presentation 1st | 2nd | 3rd | 4th |
|---|---|---|---|---|---|
| Ham | 4! | 7.71 | 6.87 | 6.21 | 6.46 |
| | cccc | 7.46 | 7.42 | 7.37 | 7.21 |
| | tttt | 7.12 | 7.21 | 6.83 | 6.96 |
| | cctt | 7.71 | 7.08 | 5.50 | 6.00 |
| | ttcc | 6.54 | 7.17 | 6.62 | 6.58 |
| Pork | 4! | 6.25 | 5.88 | 6.38 | 6.17 |
| | cccc | 6.62 | 6.54 | 6.79 | 6.88 |
| | tttt | 6.50 | 6.25 | 6.21 | 6.50 |
| | cctt | 6.58 | 6.47 | 5.75 | 6.00 |
| | ttcc | 5.75 | 6.17 | 6.17 | 5.67 |
| Chicken, W. | 4! | 7.20 | 7.04 | 6.71 | 7.25 |
| | cccc | 7.58 | 7.46 | 7.29 | 6.92 |
| | tttt | 6.88 | 7.04 | 7.00 | 6.92 |
| | cctt | 7.42 | 6.88 | 6.38 | 6.33 |
| | ttcc | 6.83 | 6.96 | 7.17 | 6.92 |
| Chicken, D. | 4! | 7.12 | 6.12 | 6.38 | 6.00 |
| | cccc | 6.83 | 6.50 | 6.71 | 6.33 |
| | tttt | 6.50 | 6.46 | 6.83 | 6.33 |
| | cctt | 6.67 | 6.75 | 5.58 | 6.04 |
| | ttcc | 6.04 | 5.75 | 6.08 | 5.58 |
| Combined | 4! | 7.07 | 6.48 | 6.42 | 6.47 |
| | cccc | 7.12 | 6.98 | 7.04 | 6.84 |
| | tttt | 6.75 | 6.74 | 6.72 | 6.68 |
| | cctt | 7.10 | 6.80 | 5.80 | 6.09 |
| | ttcc | 6.29 | 6.51 | 6.51 | 6.19 |

TABLE 5. Probability Values* for Orthogonal Comparisons of Position Effect of Mean Preference Ratings in Which Four Meat Samples Were Presented to a Subject

| | | Pairing | cc | | tt | | ct |
|---|---|---|---|---|---|---|---|
| Position | Meat / | Plan | 4 | 2 | 5 | 3 | 1 |
| 1 vs 2 | Ham | | .11 | .44 | .95 | .60 | .19 |
| | Pork | | .38 | .39 | .82 | .21 | .18 |
| | Chicken, W. | | .07 | .32 | .66 | .73 | .29 |
| | Chicken, D. | | .58 | .12 | .21 | .45 | .02 |
| | $\chi^2$ Comb'd P | | .13 | .27 | .81 | .61 | .03 |
| | | | (.12) | | (.82) | | |
| | | Plan | 5 | 2 | 4 | 3 | 1 |
| 3 vs 4 | Ham | | .44 | .28 | .83 | .64 | .70 |
| | Pork | | .14 | .62 | .75 | .82 | .25 |
| | Chicken, W. | | .22 | .08 | .44 | .38 | .96 |
| | Chicken, D. | | .09 | .09 | .87 | .08 | .22 |
| | $\chi^2$ Comb'd P | | .10 | .10 | .93 | .41 | .58 |
| | | | (.05) | | (.79) | | |
| | | Plan | | 2 | | 3 | 1 |
| 1+2 vs 3+4 | Ham | | | .22 | | .15 | .008 |
| | Pork | | | .88 | | .45 | .77 |
| | Chicken, W. | | | .013 | | .50 | .26 |
| | Chicken, D. | | | .22 | | .66 | .11 |
| | $\chi^2$ Comb'd P | | | .06 | | .48 | .03 |

*Probability Values Presented in This Table Reflect the Test of Significance Regarding the Hypotheses of $\bar{x}_1 > \bar{x}_2$ , $\bar{x}_3 > \bar{x}_4$ and $\bar{x}_{1+2} > \bar{x}_{3+4}$ .

Combined Results Were Obtained by the Chi-Square Method of Combining Results of Similar Experiments(3). Individual Probabilities in the Table are Based on an N of 24.

TABLE 6. Error Terms from Five Experimental Plans and Four Meat Products

| Plan / | Meat Products | | | | |
|---|---|---|---|---|---|
| | Ham | Pork | Chicken, W. | Chicken, D. | Composite |
| 41 | 2.7386 | 1.9697 | 1.1203 | 2.9246 | 2.1883 |
| cctt | 3.0850 | 1.6721 | 1.4882 | 1.9670 | 2.0531 |
| ttcc | 1.6356 | 2.5553 | 1.1933 | 1.5490 | 1.7333 |
| tttt | 1.6360 | 1.1497 | 0.8085 | 1.4420 | 1.2590 |
| cccc | 0.9134 | 1.0659 | 0.7993 | 0.9171 | 0.9239 |

*Bracketed numbers indicate the variances which are not significantly different at the probability level of 0.05.